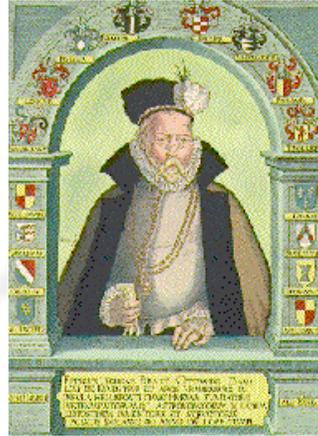


The Tycho Brahe Corpus of Historical Portuguese



<http://www.tycho.iel.unicamp.br/~tycho/corpus/>

Currently associated to the research project:

[Portuguese in time and space: Linguistic Contact, Grammar Competition and Language Change](#)

(Fapesp grant 2012/06078-9)

The Tycho Brahe Corpus of Historical Portuguese - *TBC*

- 65 texts from authors born between 1385 and 1881 ± 2.800.000 words
- annotated along the lines of the *Penn Parsed Corpora of Historical English*
(<http://www.ling.upenn.edu/hist-corpora/>)
- Part of Speech (POS) tagging (± 1.500.000 words currently available, out of 34 texts)
- Syntactic Parsing (± 700.000 words currently available, out of 16 texts) – 400.000 more in final process of revision (from newspapers, proceedings, letters, plays).

By Edition Level

Preserved orthography

Edited orthography

By Annotations

POS Annotation

Syntactic Annotation

By Place of Production/Publication

Note:
This criterion refers to the country where the text has been produced (manuscripts) or pulished (printings). For simplicity, Portugal refers to any European country.

Portugal

Brazil

Atas dos Brasileiros
Cartas Brasileiras: cultos
Cartas para Cícero Dantas Martins, Barão de Jeremoabo (1880-1903)
Cartas para vários destinatários (1809-1904)
Iracema, lenda do Ceará
Maria ou a menina roubada

Unknown/other

Created: ... de Faria
Last updated: 2015-01-27 17:43:45 (Brazil-West)

If you see any strange characters on screen, please check if the browser encoding is set to UTF-8.

What written corpora can tell us about the roots and development of Brazilian Portuguese?

- The influence of European Portuguese;
- The role of African languages (of African L2 learners of Portuguese).

Many things... thanks to a great diversity of sources of relevant data

- 16th and 17th c. Portuguese (Classical Portuguese *CIP*);
- the dynamics of change from CIP to Modern European Portuguese *EP* – from 1700 on;
- A great diversity of sources for historical Brazilian Portuguese *BP*: recent work by several research groups in Brazil has produced a great quantity of documents and the aim of the *TBC* is to establish partnership with people editing documents of every possible kind, using for them the same tools of annotation that have been used with literary - and/or representative of more standard language - texts, and addressing to them the same questions, with the same automatic searches.



THE MORPHOSYNTACTIC INNOVATIONS OF BP

A summary

- Topic-oriented language properties (Pontes 1981; Galves, 1998; Negrão, 1999; Negrão e Viotti 2000; Kato and Duarte 2008; Avelar and Cyrino, 2008; Negrão e Viotti, 2008; Avelar 2009; Avelar and Galves, 2011, Naves and Pilati, 2012; Andrade and Galves 2014);
- Impoverishment of inflectional morphology (Guy 1981, Scherre 1988, Holm 2004);
- Partial null subjects (Figueiredo Silva 1996; Kato and Negrão, eds. 2000; Rodrigues 2000, 2002, 2004; Holmberg et al. 2010);
- Pronominal system in which null objects/1st and 2nd person clitics/non-case marked tonic pronouns co-exist (Galves 2000, a.o);
- Generalized proclitic placement to the main verb (Pagotto 1992; Galves et al. 2005);
- Null singular determiners with generic interpretation (Muller 2002; Schmitt and Munn 2003; Muller and Oliveira 2004; Pires de Oliveira and Rothstein 2011; Cyrino and Espinal 2014) .

Topic-oriented language properties (locative/genitive subjects)

1. a) *aparece* imagem **de/em apenas 3 desses cinco monitores**

appear-3SG image of/in only 3 of-those five monitors

b) **apenas 3 desses cinco monitores** *aparecem* imagem
only 3 of-those five monitors appear-3PL image

‘only the image of three out of five monitors appears’
(Avelar & Galves 2014)

Locative and genitive subjects in Bantu languages

KINANDE (Baker 2003: exemplo 25)

Omo-mulongo mw-a-hik-a (?o-)mu-kali
LOC.18-village **18S**-T-arrive-FV (AUG)-CL1-woman.1
'At the village arrived a woman'

LUBUKUSU (Diercks 2011: 703)

Mú-mú-siirú mw-á-kwá-mó kú-mú-saala
18-3-forest **18s**-PST-fall-**18L** 3-3-tree
'In the forest fell a tree'

ZULU (Buell 2004: 3)

I-sikole si-fund-el-a a-bantwana
7-7.school **7.SBJ**-study-APPL-FV 2-2.child
(Lit. 'The school studies at the children')

CHICHEWA (Simango 2007: exemplo 23)

Mavuto a-na-f-a maso
Mavuto SM-PST-die-FV eyes
'Mavuto became blind' (Lit. 'Mavuto died eyes')

Avelar and Galves 2014

Impoverishment of inflectional morphology

2. Os menino(s) chegou/chegaram
the-PL boy-(PL) arrived-SG/arrived-PL
'The boys arrived'
3. Um dos mais velho orixás
one of-the.PL more old.SG deities
'one of the more ancient deities'

Partial null subjects

- Tendency for the subject position to be lexically filled (Duarte, 1995 and subsequent work);
- Null subjects are not allowed in all contexts (Figueiredo Silva 1996; Ferreira, 2000; Rodrigues, 2004)
- Singular null subjects can receive an arbitrary interpretation (Holmberg et al. 2010).

Arbitrary null subjects and null objects: a highly ambiguous structure

4. C: Onde está **minha cueca com personagens de desenhos animados?**

‘Where is my underwear with characters of cartoons’

M: Está lavando, Calvin

is washing

[pro_{arb} está lavando pro_{TOP}] ‘Somebody is washing it’

[pro_{TOP} está lavando] ‘It is being washed’



Pronominal paradigm and placement

5. a) **te** vi ontem
CL2 saw yesterday

b) vi **você** ontem
saw you yesterday

'I saw you yesterday'

6. a) Eu não tinha **te** visto ontem (EP: **te** tinha)
I not had CL2 seen yesterday

b) Eu não tinha visto **você** ontem
I not had seen you yesterday

'I had not seen you yesterday'

Portuguese revisions on Paulo Coelho's novel O Alquimista

(7) a. **Me** chamo Fátima – disse a moça... (BP)

CL1 call Fatima said the girl

b. Chamo-**me** Fátima – disse a moça (EP)

call –CL1 Fatima said the girl

‘My name is Fatima, said the girl’

(8) a. Como **tinha** **se** **comportado** de maneira correta... (BP)

As had CL3REFL behaved in manner correct

b. Como **se** **tinha** **comportado** de maneira correcta...(EP)

As CL3REFL had behaved in manner correct

‘As he had behaved correctly...’

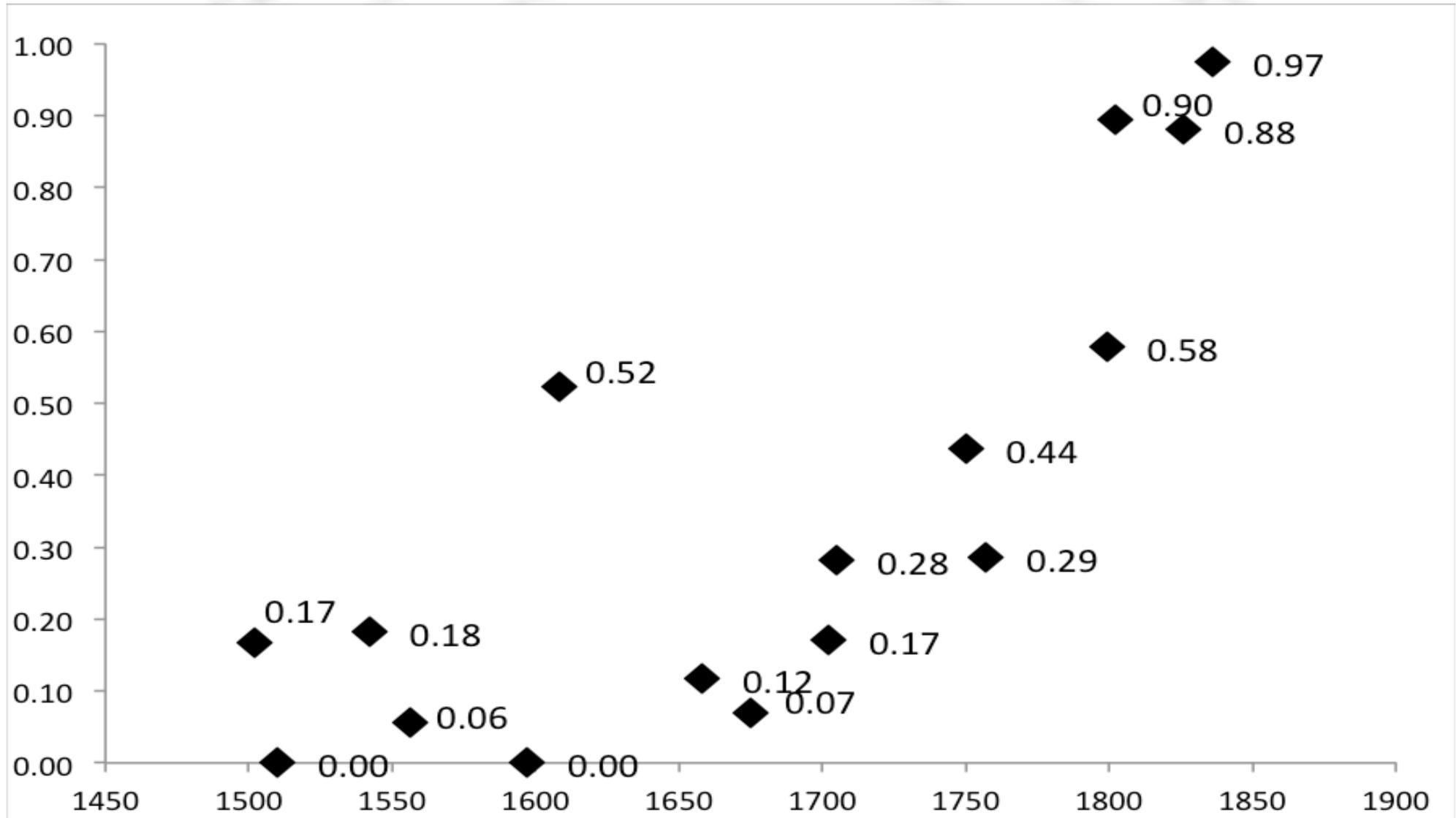
Null determiners

9. a. Crianças lêem revistinhas.
Children read-3pl comic books
Children read comic books.
b. Criança lê revistinha.
Child read-3sg comic book
'Children read comic books.' (from Schmitt & Munn 2003)
10. a. Sapos são anfíbios.
Frogs are amphibians
b. Sapo é anfíbio.
Frog is amphibian
'Frogs are amphibians.'
11. a. Pintassilgos são raros em São Paulo.
goldfinches are rare in São Paulo
b. Pintassilgo é raro em São Paulo.
goldfinch is rare in São Paulo
'Goldfinches are rare in São Paulo.'

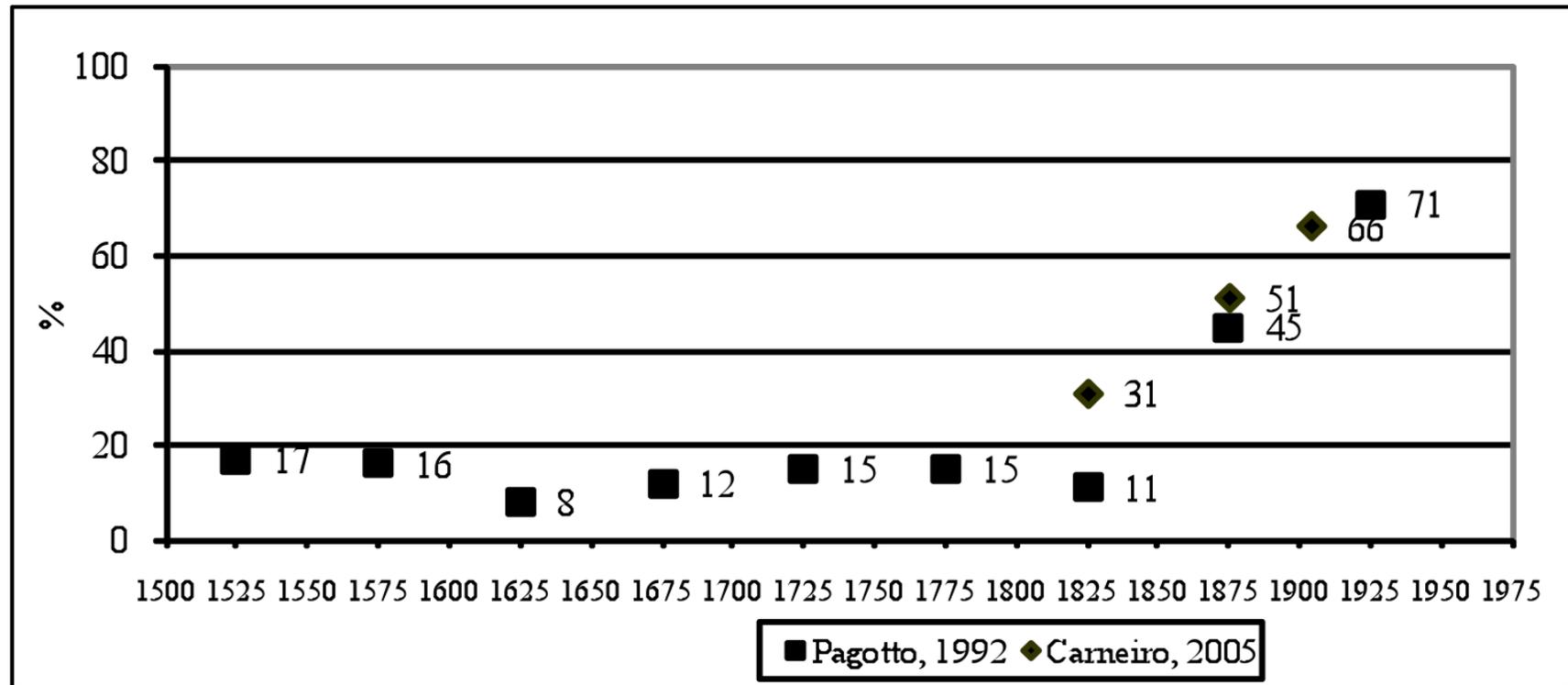


THE INFLUENCE OF EUROPEAN PORTUGUESE

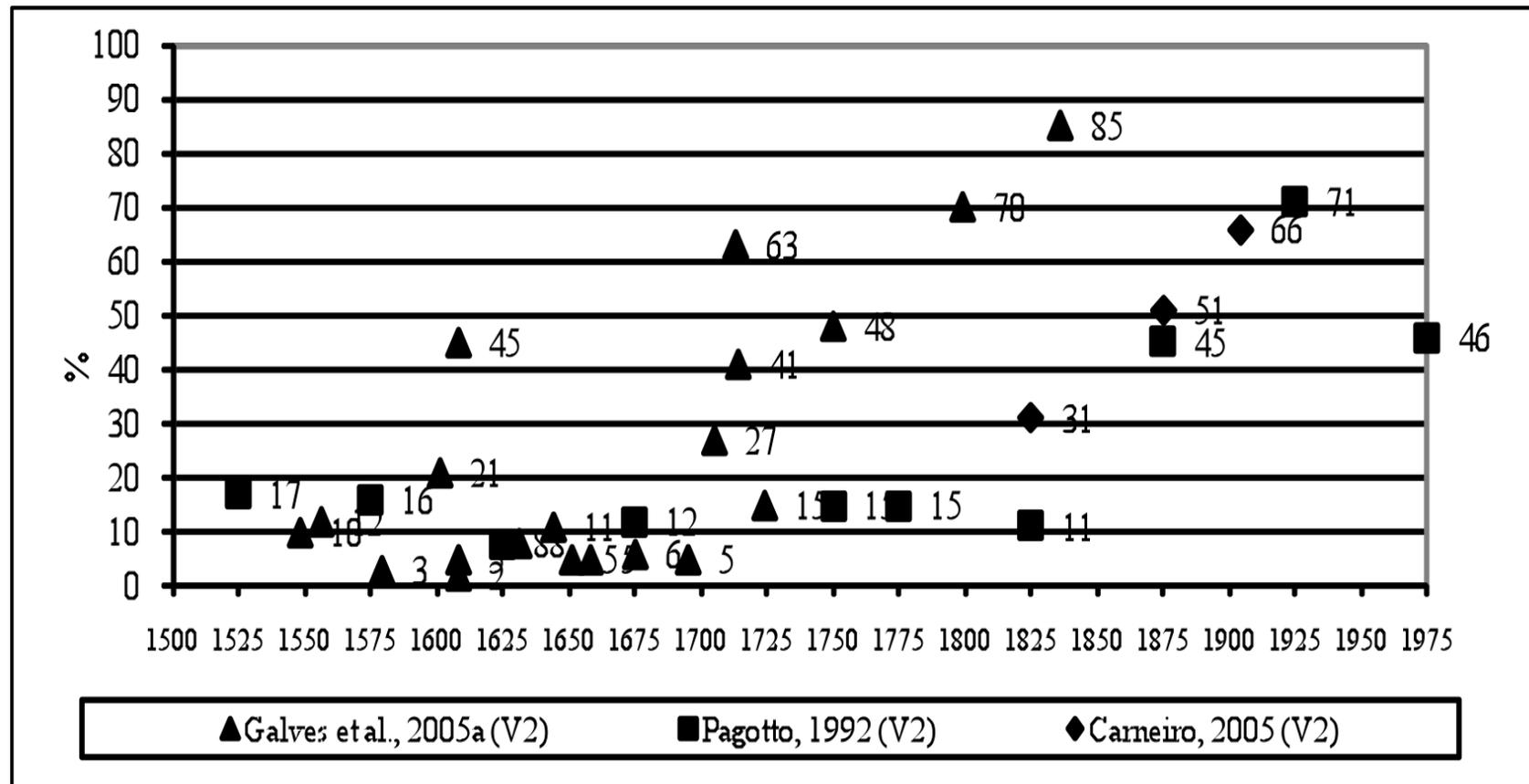
Evolution of enclisis/proclisis in clauses with pre-verbal subjects (SV-cl / S-clV) - CIP to EP (Galves & Paixão deSousa 2013)



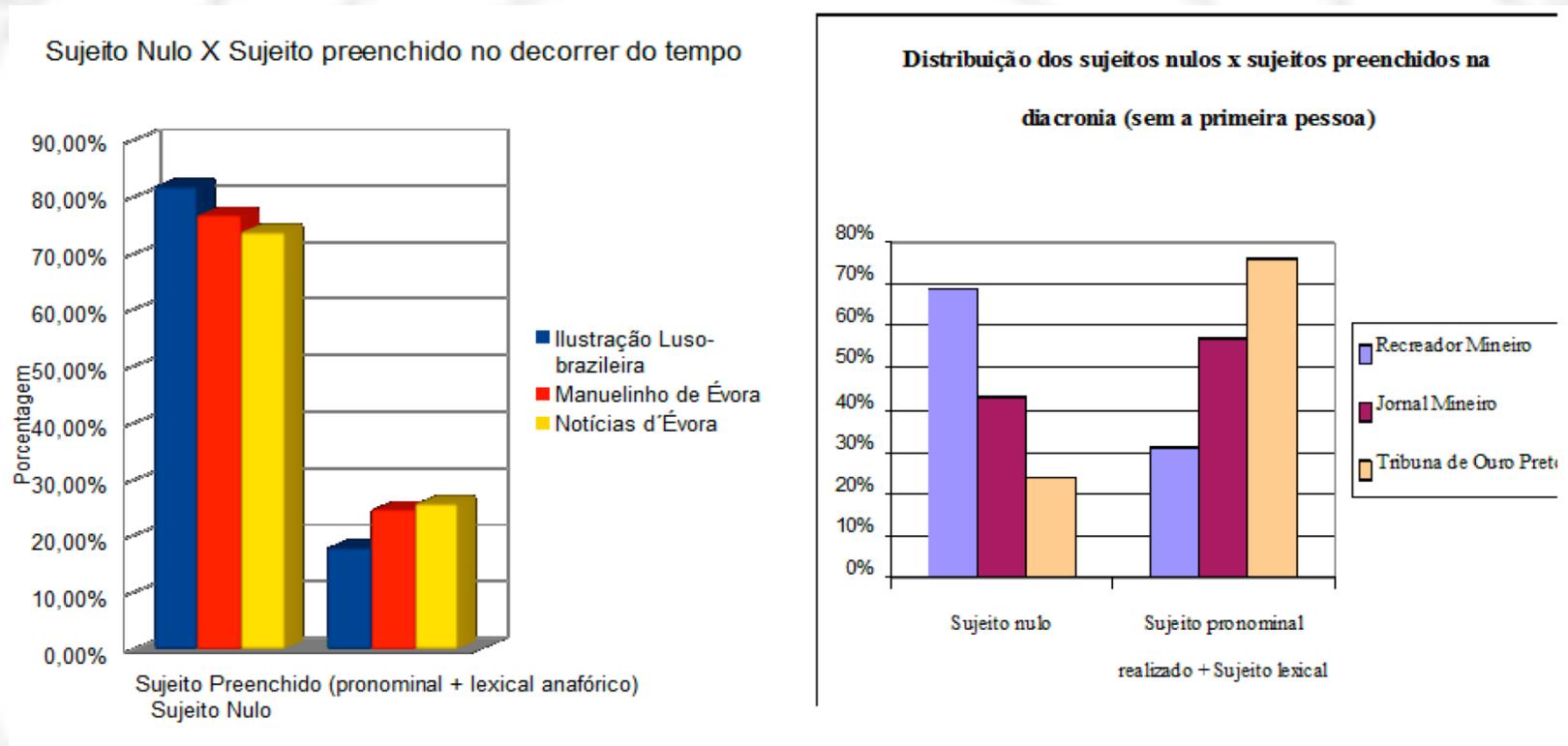
The evolution of clitic-placement in V2 non proclitic contexts (XV-cl vs. XclV) **CIP to BP**



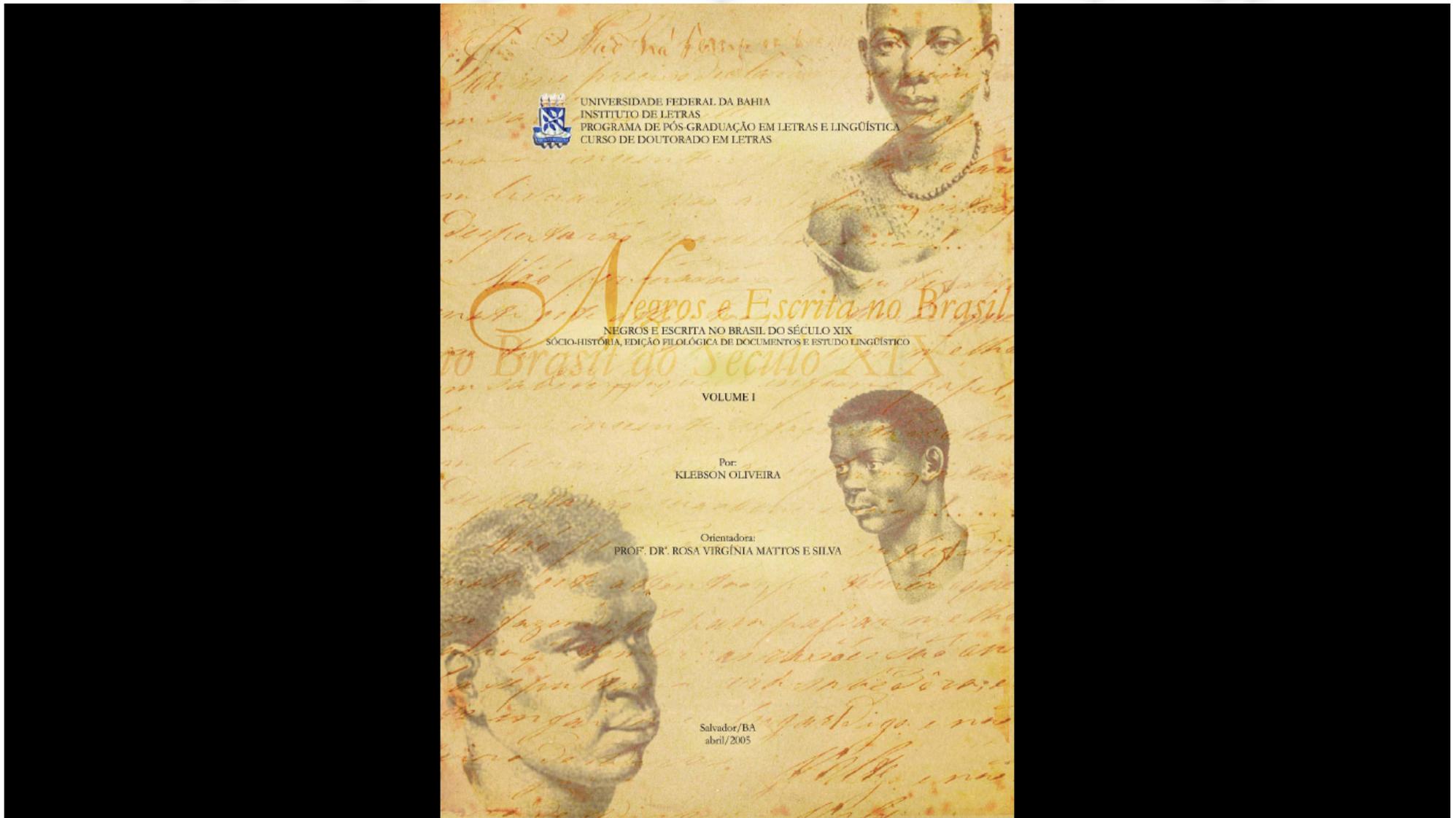
CLP to EP (black Δ) and CLP to BP (■◆): a parallel evolution *in texts* until the first half of the 20th c.



The evolution of the use of null subjects in newspapers 1850-1950 – EP vs BP (Gravina 2013)



Closer to Afro-Brazilian speakers: the proceedings of the *Sociedade Protetora dos desvalidos*



Proceedings of the *Sociedade Protetora dos Desvalidos* (Oliveira 2003, 2006)

233 documents written from 1837 to 1894, by 23 authors (11 'certainly born in Brazil', 12 'probably born in Brazil'). The data are drawn from the 'certainly born in Brazil' authors .

They instantiate a very peculiar grammar competition, resulting from a clash between oral and written competence.

The imperfect literacy of the writers gives us very interesting hints about the difficulty of processing functional words, given their peculiar morphology in Portuguese.

Agreement facts

- 12. Não pode esta Vossa Senhoria ciente de todos os
- NEG can be Your Lordship aware of all the.PL
- trabalho da casa
- work.SG of-the house
- ‘ You cannot be aware of all the works of the house’
- 13. pedi palavra algum senhores sócios
- ask.SG word some.SG Sir.PL member.PL
- ‘Some members ask permission to speak.’
- 14. Foram eleito os irmãos seguinte
- were.PL elected.SG the.PL brother.PL following.SG
- ‘The following brothers were elected.’

Null subjects: contradicting properties

- Arbitrary null subjects:

15. E se não for bom reúne
and if NEG be good meet
'and if it is not good, we meet'

- Null subjects in contexts that are OK in ClP/EP, and * in modern BP (Figueiredo Silva 1994):

16. Foram aprovados nesta sessão os projetos [que *pro* já mencionei]
were approved in-this session the projects that (I) already mentioned

17. Pede a palavra ... Theodoro Gomes, para pedir a leitura do seu
asks the word Theodoro Gomes, to ask the reading of his
requerimento, [que *pro* mandou a mesa na sessão passada,]
request which (he) sent to the board in-the last session

Null determiners

18. Depois de lido o senhor presidente disse que **assembléia** decidisse
after of read the Sir president said that Assembly decide
'After reading it, the president said that the Assembly had to take a decision'
19. Mando O Senhor Prizidente **ler, acta, do dia 6 de Marco**
ordered the Sir President read proceeding of-the day 6 of March
'The president ordered that the proceeding of the 6th of March were read'

Segmentation facts:

20. nada dicerao Os Irmãos **da Semblea** atal respeito
nothing said the brothers of-the assembly at this respect
'The brothers of the Assembly said nothing at this respect'
21. **a prezentou** huma Planta oferecida pello *Ilustríssimo* *Senhor* Jozé Corrêa
(CL?) presented a plan offered by-the most-illustrious Sir Jozé Correa
'he presented a plan offered by Sir Jozé Correa'
22. Eu **a signei** eu **asignei**
I CL (?) signed I CL(?) signed

Unexpected cases of lack of agreement possibly due to the phonetic context (before a fricative)

22. O senhores socios
 the.SG Sir.PL member.PL

23. O artigo 45 do **Estatutos**
 the article 45 of-the.SG statute.PL

24. O **juros**
 the.SG interest.PL

The Tycho Brahe Corpus of Historical Portuguese



Thank you!

<http://www.tycho.iel.unicamp.br/~tycho/corpus>

TBC workflow

1. XML electronic editing using the ***eDictor tool***;
2. Automatic POS tagging (with Fabio Kepler's tagger) incorporated to eDictor;
3. Intermediate revision of the edition and POS tagging;
4. Automatic syntactic annotation (with ***Dan Bickel's parser***);
5. Semi-automatic revision of the syntactic annotation (with ***Corpus Draw*** or ***Annotald*** Annotald – <http://annotald.github.com/user.html>);

e-dictor

eDictor is a tool for philological edition and automatic linguistic annotation (cf. <http://edictor.net>)
It was developed by Maria Clara Paixão de Sousa (USP), Fábio Kepler (UniPampa) and Pablo Faria (Fapesp/Unicamp)

Projects currently using eDictor

- [Corpus Anotado do Português Tycho Brahe](#)
(Universidade Estadual de Campinas)
- [Grupo de Pesquisas Humanidades Digitais](#)
(Universidade de São Paulo)
- [Laboratório de História do Português Brasileiro](#)
(Universidade Federal do Rio de Janeiro)
- [P.S. – Projeto para o Arquivo Digital de Escrita Quotidiana em Portugal e Espanha na Época Moderna](#)
(Universidade de Lisboa)
- [Corpus Eletrônico de Documentos Históricos do Sertão, CEDOHS](#)
(Universidade Estadual de Feira de Santana)
- Memória Conquistense
(Universidade Estadual do Sudoeste da Bahia)
- Projeto do Libolo
(Universidade de Macau e Universidade de São Paulo)



Transcrição Edição Morfologia

Propriedades

Tipo:

Língua: N I S

Aplicar alterações

Substituir similares

Edição

Original: Análise fonológica Capitalizar

Edição: >>

Forma fonológica:

Lista de edições

que estes | Padronização
qu' estes | Modernização
qu estes | Grafia
qu eltes | Segmentacao

Remover Remover tudo

Texto:

[header]

#LIV

[section]+

[<p:chorus_start] [s] [...] ensinai-me por que via \
passarei à_salvação .

[p:chorus_inner] [s:character] Anjo . [s] Conhecias tu a Deus ?

[p:chorus_inner] [s:character] Moça . [s] Muito bem . [s] Era redondo .

[p:chorus_inner] [s:character] Anjo . [s] Esse era o mesmo dos céus .

[p:chorus_end] [s:character] Moça . [s] Mais alvinho que estes véus \
o vi eu vezes avondo .

[p:chorus_start] [s] Como o sino começava \
logo deitava a correr .

[p:chorus_inner] [s:character] Anjo . [s] Que lhe dizias ?

[p:chorus_inner] [s:character] Moça . [s:verse_indent1] Folgava \
e toda me gloriava \
em ouvir missa e o ver .

[p:chorus_inner] [s:character] Anjo . [s] Pastora bom era isso .

[p:chorus_end] [s:character] Diabo . [s] Era a maior mexeriqueira \
gulosa , que de_improviso \
se não andavam sobre aviso \
lá ia a ceia e a cepeira .

As 10 1/2 da Noiti não tendo mais o q trata o Prizidente suspendeu a çção

Eu 2º. Secretario q fiz i a Singno esta Acta Julio Capitolino da Boa Morte

DOCUMENTO 01

Escrito por Manuel Anastácio **Cajueiro** em 06 de outubro de 1867.

Secaõ do dia 6 de bro de 1867
lançada Seccaõ 6 bro 1867

Prizidencia Sr Manoel Salustiano. Abrio a secaõ ameio dia feita a xhamada na- forma do custume estava prizente 17 Sr Socio lida atta exterior foi aprovada Sr Pri- zidente mandou lei peilo 1 Secretario usis- tatuto de pois de lido Sr Prizidente levo- u ocuisimento da Sembreira um riqi- rimento do Sr Manoel Leonardo. Depois de lido o dito riqirimento Sr Prizidente dissé está idiscução. Pidio aparavra idis- Socio Andre xhavier idisse enquanto amim eu axho bom sifazer ascaneiras Pidio aparavra o Socio Francisco Ancielmo idisse enquanto eu axho bom sifazer mas com q dinhairo sifazer ellas. quero q vosa Senhoria diga com udinheiro da- Lotaria eu não axho bom sifazer com- elle pos u dinheiro hé para a caza de azilho não hé para outra coza. Pidio aparavra o Socio Manoel Leonardo i dis- sé com udinheiro da Lotaria q si pode sifa- zer q está vencendo poco júro isifazendo acaneira si rendes más. Pidio apara- vra Socio Narcizo idisse eu axho bom sifazer priziza tomar uma midi- da para entaõ sifazer por q udinhei- ro he para a caza de a zilho. por q en outras suedade tem sifeito por Cotas do Socios então eu não axho

Firefox Arquivo Editar Exibir Histórico Favoritos Ferramentas Janela Ajuda [va_002]: Versão para leitura [va_002]: Versão para leitura [va_002]: Versão para leitura Nouvel onglet

www.tycho.iel.unicamp.br/~tycho/corpus/cgi-bin/getversion_edictor.pl

Cajueiro SEARCH vuze remote Amazon eBay Facebook Twitter Vuze Blog [50]

Francisco de Paula da Silva Bruno Estaquio do Sacramento aprovado sendo a jóia descontado da mesma quantia depositada Foi lido o requerimento do senhor Fransisco de Paula da Silva Aprovado Às 10½ da noite não tendo mais o que tratar o presidente suspendeu a sessão

Eu 2º secretário que fiz e assino esta ata Julio Capitolino da Boa Morte

DOCUMENTO 01

Escrito por Manuel Anastácio **Cajueiro** em 06 de outubro de 1867.

Sessão do dia 6 de outubro de 1867 lançada Sessão 6 outubro 1867 Presidência senhor Manoel Salustiano. Abriu a sessão a meio-dia feita a chamada na forma do costume estavam presente 17 senhor sócio lida ata anterior foi aprovada Senhor presidente mandou ler pelo 1 secretário os estatuto Depois de lido senhor presidente levou ao conhecimento da assembleia um requerimento do senhor Manoel Leonardo. Depois de lido o dito requerimento senhor presidente disse está em discussão. Pedeu a palavra e disse sócio André Xavier e disse enquanto a mim eu acho bom se fazer ascaneiras Pedeu a palavra o sócio Francisco Ancielmo e disse enquanto eu acho bom se fazer mas com que dinheiro se fazer elas. Quero que Vossa Senhoria diga com o dinheiro da Loteria eu não acho bom se fazer com ele pois o dinheiro é para a casa de asilo não é para outra coisa. Pedeu a palavra o sócio Manoel Leonardo e disse com o dinheiro da Loteria que se pode se fazer que está vencendo pouco juro e se fazendo acaneira se rende mais. Pedeu a palavra sócio Narcizo e disse eu acho bom se fazer precisa tomar uma medida para então se fazer porque o dinheiro é para a casa de asilo. Porque em outras Sociedade tem se feito por cotas dos sócios então eu não acho

Cajueiro Atingido o início da página, continuando do fim Realçar tudo Diferenciar maiúsculas/minúsculas

BP nominal agreement variation according to socio-linguistic factors

- Afro-Brazilian isolated communities, 16%
- illiterate fishermen 38 %
- illiterate people from the city of Rio de Janeiro 48%
- educated people from the city of Rio de Janeiro 73%
- people from the city of Rio de Janeiro with university degrees 94%

(Lucchesi et al., 2009: 348)

Writers of the Atas (first approximation concerning D and N)

191/916 79%